Supercharged Digital Collections:

Moving to the fast lane with scalable open source infrastructure

J. Willard Marriott Library
THE UNIVERSITY OF UTAH

ALL U NEED

Photos from University of Utah, Multimedia Archives Photo Collection, Gary Dean Brown

Presenters

Anna Neatrour

Brian McBride

Matt Brunsvik

Harish Maringanti

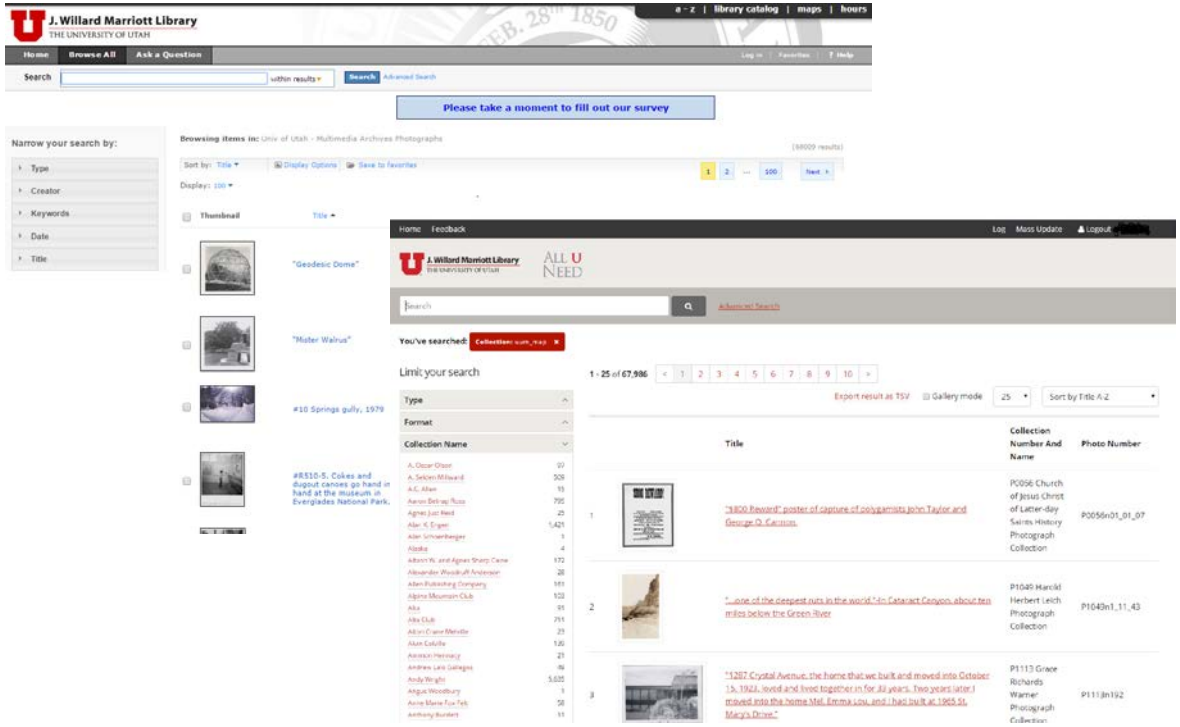
Other Presenters/contributors:

Jeremy Myntti

Alan Witkowski

What we'll cover today

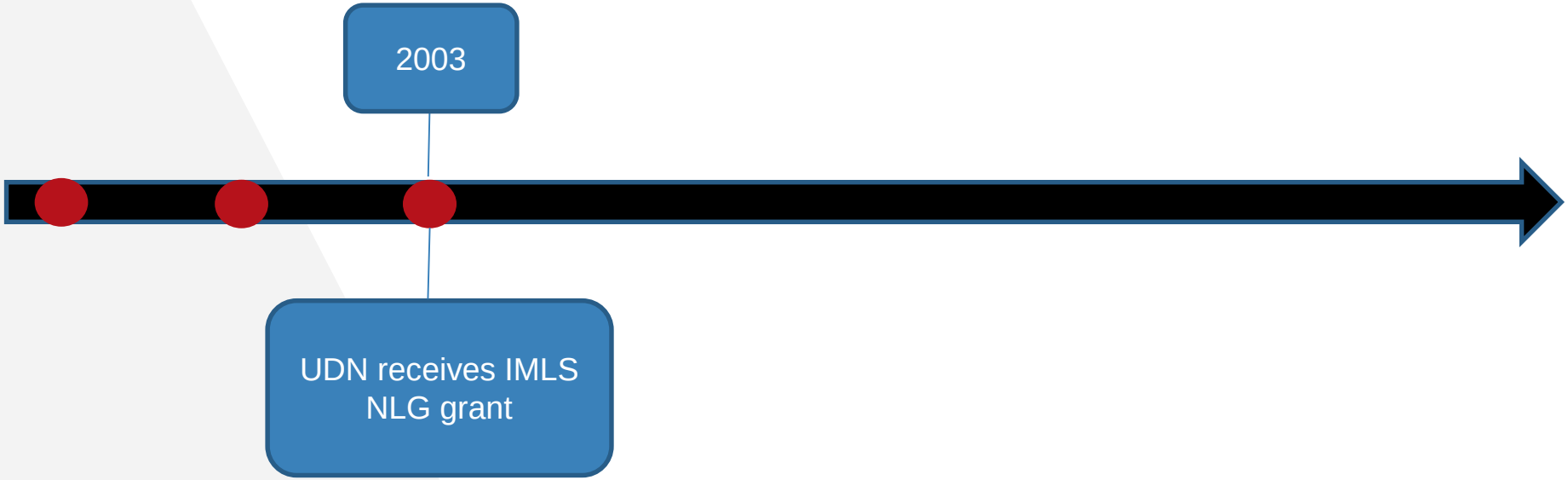
- History
- SIMP demo
- Migration
- Lessons learned



History of UU Marriott Digital Library

History of UU Marriott Digital Library

History of UU Marriott Digital Library



History of UU Marriott Digital Library

2010

UDN reaches 1 Million pages milestone

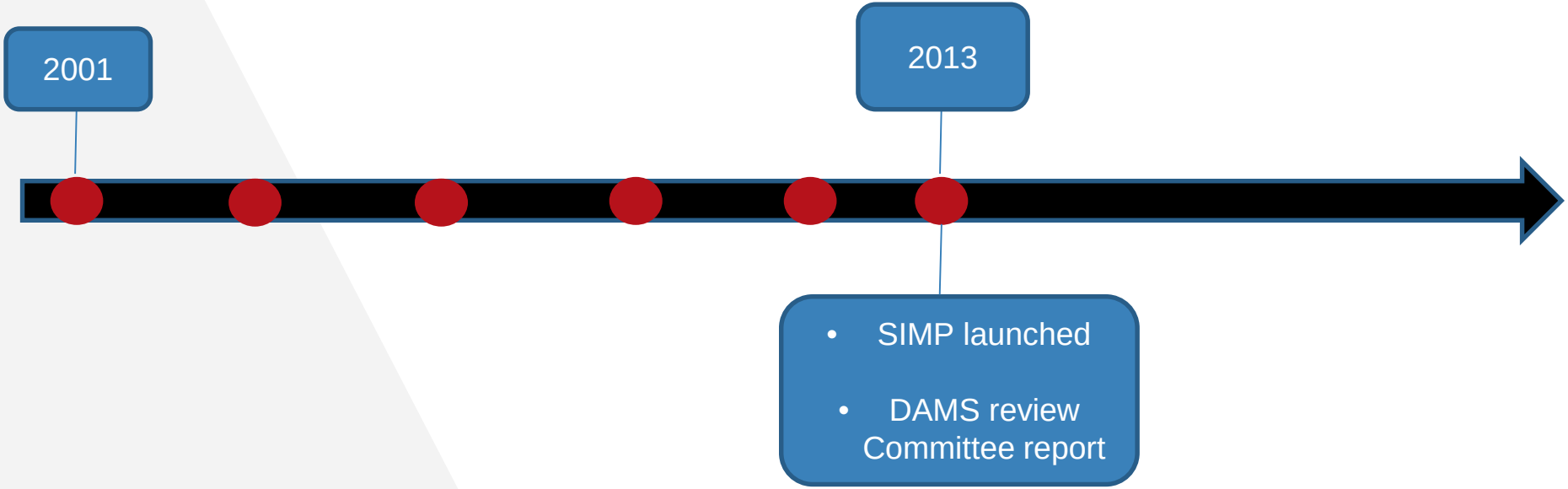
History of UU Marriott Digital Library

2001

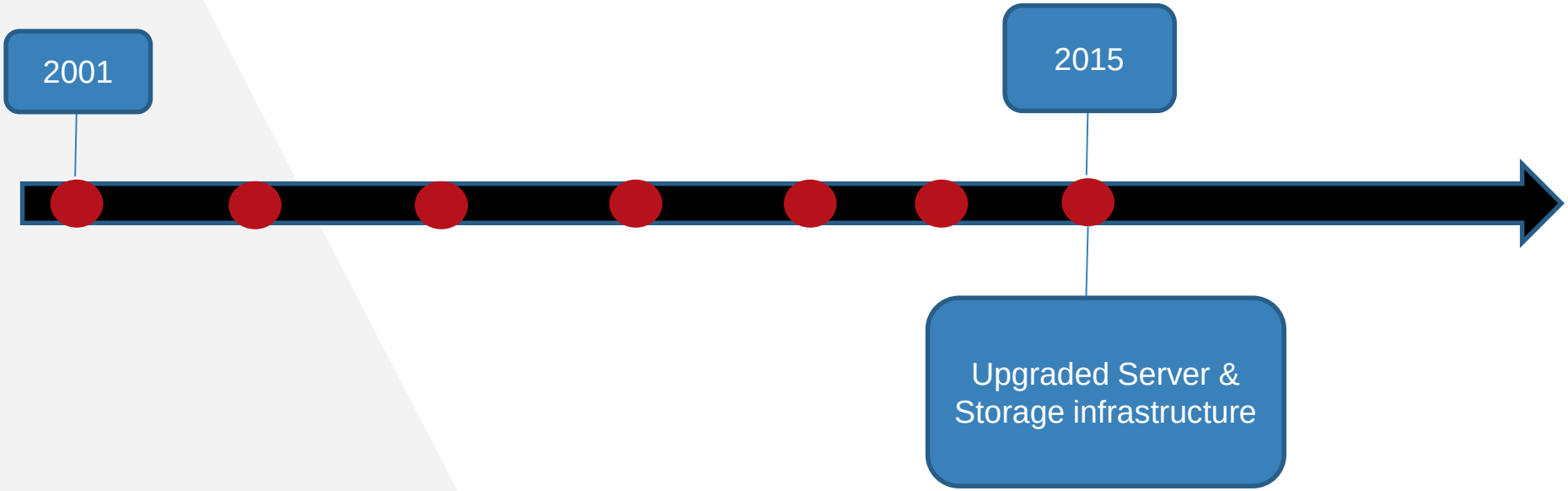
2012

- Purchased Rosetta for long term archiving
- SIMP development started

History of UU Marriott Digital Library



History of UU Marriott Digital Library



History of UU Marriott Digital Library

2001

2016

UDN migrated (Jan – June)

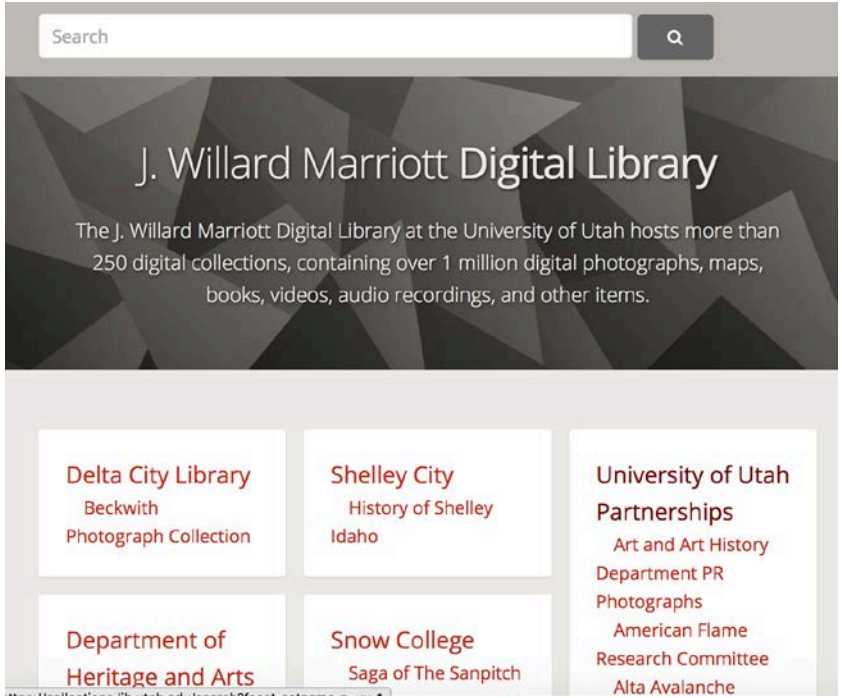
All Other digital Collections migrated (June – Dec)

Marriott Library Digital Collections

- 269 collections
- 717,858 IEs (items)
- 139,583 compound objects
- 12 file types

(images,videos,pdfs...)

- 3 TB storage
- Over 50 partners (internal and external) hosting collections with the Marriott Library



https://collections.lib.utah.edu/

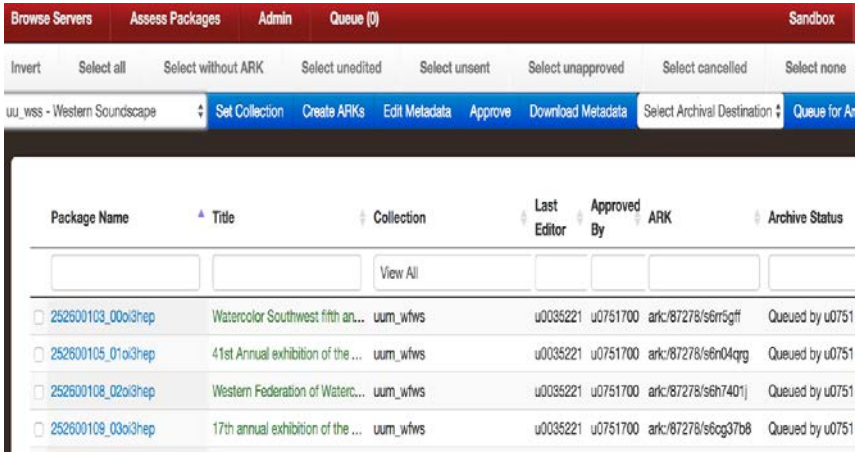
Utah Digital Newspapers (UDN)

- ~1.926 Million pages
- ~21.64 Million articles
- ~138 Newspaper titles
- Article level newspaper data
- Many unique papers not in Chronicling America

https://digitalnewspapers.org/

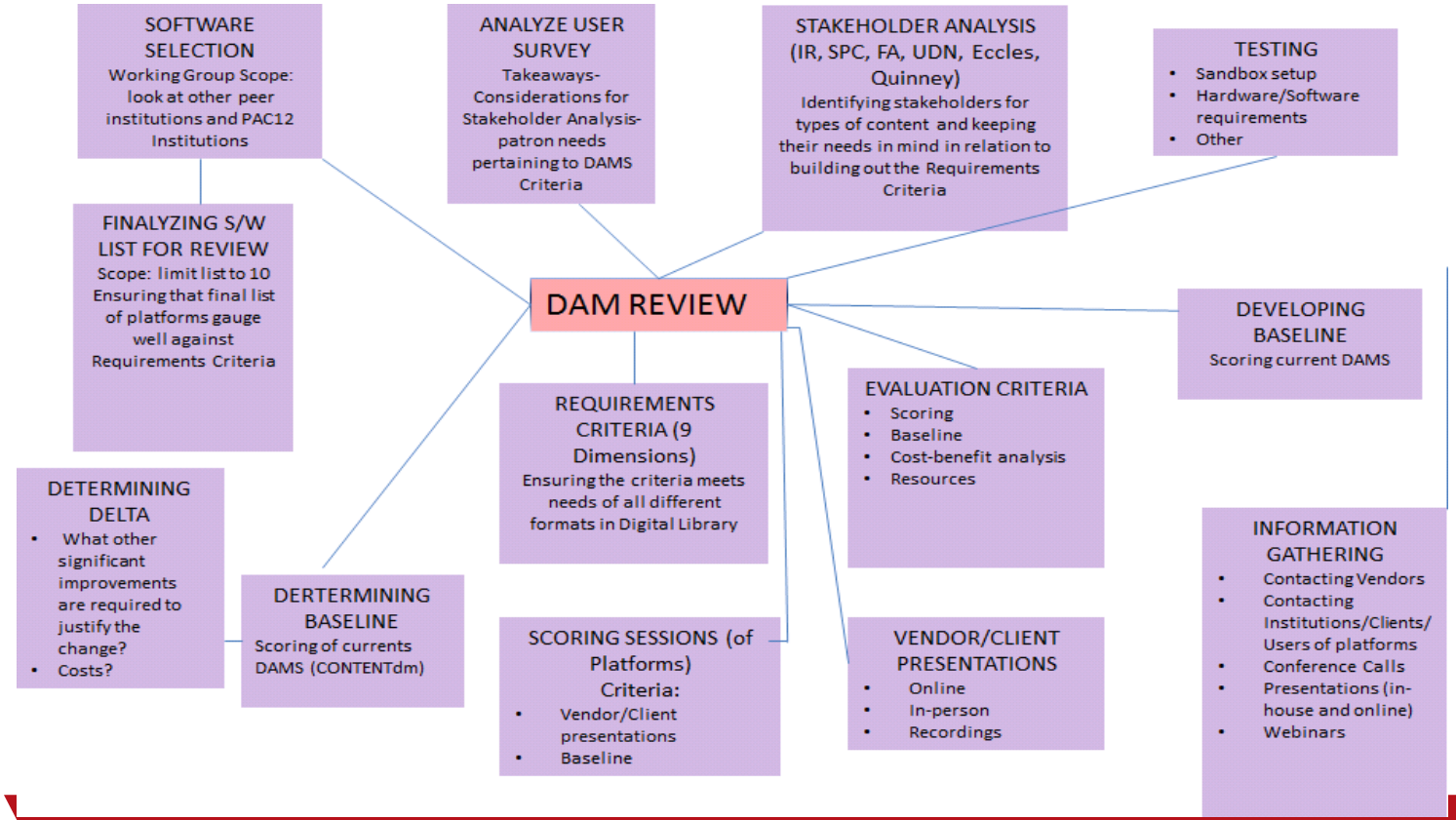
SIMP Tool for metadata management

- Initially designed to facilitate management between CONTENTdm and Rosetta, our preservation system
- Platform agnostic and modular, provides ability to use other DAMs/Preservation Systems
- Options for automatically extracting some metadata (format, OCR text)
- Updated in 2016 to support Solphal



More details available in D-Lib article:
http://www.dlib.org/dlib/july14/neatrour/07neatrour.html

DAMS Review in 2013



DAMS Review in 2013

DAMS Review - requirements criteria

Patron ease of use
•End-user experience
•Search options and accuracy - discoverability of the content
•Handle all formats in similar fashion for all browsers
•ADA compliance

Types of content supported
•Major content categories (e.g. IR, newspapers, EAD)
•File types and datasets
•Tiers of content (both object and item level)
•Media streaming

Issues related to ingest/conversion/exit
•Workflows for adding new content
•Converting existing collection data
•Barrier(s) to exit

Collection administration
•Interface design
•User permissions
•Website configuration
•Statistics/Reporting/Logs

Metadata administration
•Metadata creation, editing, and maintenance options
•Faceting
•Copyright/Embargo
•Search Engine Optimization (SEO)

Integration with other library platforms
•Discovery layer (e.g. Primo)
•Digital preservation system (e.g. Rosetta)
•Locally developed tools (e.g. SIMP Tool)

Technical infrastructure/administration
•Scalability of collections/content
•Hardware/Server(s) specs
•API support
•System and server installation, configuration, and upgrades
•Internal resources required to run the system

Technical support
•Training
•Help desk

Future and strategic directions

DAMS Review Outcome - Solphal

Hardware Requirements - CONTENTdm vs Solphal

CONTENTdm

•2 servers (UDN & Digital Collections)

•Digital Collections (2x12 Core 3.0 Ghz, 64 Gb mem) UDN (2x8 Core 2.4 Ghz, 96 Gb mem)

•Scalability issues

•Limited customizations

Solphal

•3 VMs (indexing services, UDN, and Digital Collections)

•VMs avg. 2 cores, 3 Gb mem

•Extremely small footprint (hardware and power savings)

•Highly scalable and sustainable

•Fully customizable

•Easy replication and cloud-centric

Benefits of Solphal

•In-house expertise and support

•Student, patron, and partner convenience
- faster response times
- advanced search functions
- front-end metadata editing
- customized interface

•Substantial power, hardware, and licensing cost savings

•Improved workflows and QA/QC for all teams

•Staff time savings = ~1 less FTE supporting servers

•Demonstrating and sharing open source solutions with other institutions

Indexing Improvements

Utah Digital Newspapers

•55 GB of metadata

•21,648,462 records

•CONTENTdm
- Full re-index ~ 1440 hrs (site is online)
- Full re-index ~ 240 hrs (site is offline)

•Apache Solr indexing
- Initial index ~ 2.75 hrs
- Incremental index ~ instant

Digital Collections

•6 GB of metadata

•2,248,922 records

•CONTENTdm
- Full re-index ~ 144 hrs (site is online)
- Full re-index ~ 72 hrs (site is offline)

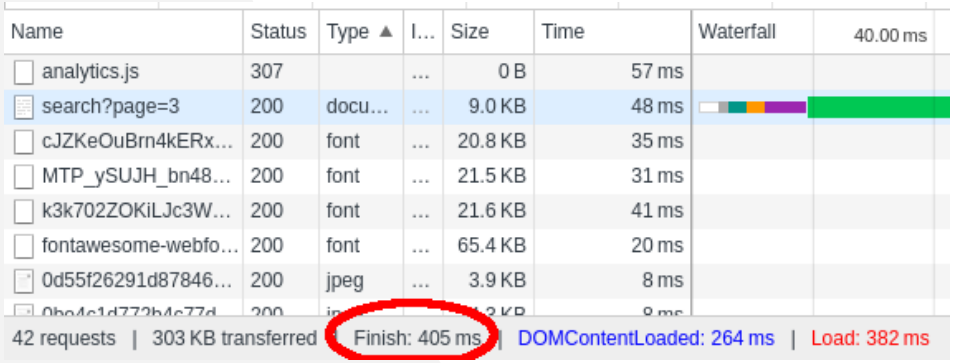
•Apache Solr indexing
- Initial index ~ 25 min
- Incremental index ~ instant

User Interface Improvements

Utah Digital Newspapers

•CONTENTdm
- Page load times (avg) 10s
- Pdf display issues

•Solphal
- Page load times (avg) under 500ms



Digital Collections

•CONTENTdm
- Page load times (avg) 5s
- Pdf display issues

•Solphal
- Custom HTML5 image viewer
- HTML response times (avg) under 50ms
- Complete page load times (avg) under 500ms

Sample Solr Ingest File

```
<add>
<doc>
<field name="title_t">U Archives D. Art 1890-1939 fd 1
#2</field>
<field name="form_t">photograph</field>
<field name="source_t">University of Utah Archives</field>
<field name="work_id_t">101</field>
<field name="date_digital_t">4/20/2010</field>
<field name="rights_t">Digital Image Copyright 2010,
University of Utah. All Rights Reserved.</field>
<field name="type_t">Image</field>
<field name="imaging_technician_t">Amelia Walchli</field>
<field name="metadata_cataloger_t">Amelia Walchli</field>
<field name="file_name_t">101.tif</field>
<field name="ark_t">ark:/87278/s6sb7cp8</field>
<field name="archival_file_t">Volume1/U Archives D. Art
1890-1939 fd 1 #2.tif</field>
<field name="created_tdt">2010-04-19T00:00:00Z</field>
<field name="modified_tdt">2010-04-26T00:00:00Z</field>
<field name="oldid_t">Art2 0</field>
<field name="id">15729</field>
<field name="parent_i">0</field>
<field
name="thumb_s">/68/84/6884a28421c31d96d4d9ec466f47fecbb99af
700.jpg</field>
<field
name="file_s">/5c/bc/5cbc0793020d24aba7eb6e4443d0c6574970ac
57.jpg</field>
<field name="format_t">image/jpeg</field>
<field name="setname_s">uu_aah_artpr</field>
<field name="restricted_i">0</field>
</doc>
<doc>
<field name="title_t">U Archives D. Art 1890-1939 fd 1
#4</field>
<field name="form_t">photograph</field>
<field name="source_t">University of Utah Archives</field>
<field name="work_id_t">102</field>
```

UDN migration vs. Digital Collections migration

Utah Digital Newspapers

•Newspapers metadata more standardized, but not perfect

•Inconsistencies with titles, dates, paper names, and type fields

•Clean-up for nonstandard characters, newspaper names issues

•Previous architecture had one newspaper title split into multiple collections (scalability issues)

•Consistent field names and metadata meant that migration happened sooner and faster

•No external partners to contact

Digital Collections

•More communication needed with both internal and external partners during migration

•Problem or missing metadata values in older collections not fixed (yet)

•Field label standardization

•Standardization of type and format

•Lots of clean-up work still to do in the future

Metadata issues: Corrections

Metadata Standardization

Establish core fields needed for faceting and queries in new DAM, and change varying fields to the following:

- Creator
- Date
- Subject
- Rights
- Spatial Coverage
- Type
- Title

Title	Mr. J. Holman Waters: Recollections of Samuel Newhouse, Salt Lake City, Utah: an interview by Hynda Rudd, July 16, 1977
Creator	Waters, J. Holman, 1912-1995
Date	1977-07-16
Spatial Coverage	Salt Lake City, Salt Lake County, Utah, United States
Subject	Newhouse, Samuel, 1853-1930; Businessmen--Utah--Salt Lake City--Interviews; Jews--Utah--Salt Lake City
Type	Text
Rights	Digital Image © 2016 University of Utah. All Rights Reserved. Contact the J. Willard Marriott Library Manuscripts Division for use and restrictions. 801-585-3082

Metadata fields gone wild!

About the Art
About the Artist
Abstract
Accession number
Accession Number
Access restrictions
Access Restrictions
Access rights
Accompanying Material
Acoustic Zone
Added Title
Added title 2
Additional information
Additional Information
Additional Resources
Additonal notes
Address
Affiliation
Ages
Alt # 1
Alt # 2
Alternate Name(s)
Alternate title
AlternateTitle
Alternate Title
Alternative formats available
Alternative Formats Available
Alternative title
Amendments and updates
Anatomy
Annotation
Archival file
Archival Resolution
ark
ARK
Article Title
Article Title(s)
Artist
Artist/Manufacturer
Artist's Notation
Athlete's Name
Audience

Audio Clip Instructions
Author
Author/Editor
Author/Editor Bio
Author(s)
AV Item Number
Awards
A-Z assignment
Band
Banner
Batch
Batch #
Batch and Box #
Batch number
bibliographicCitation
Bibliographic Citation
Binding
Birth Date
Birthplace
Bit depth
Bit-depth
Bit Depth
Box and Folder Number
browse order
Building
Business/building names
B/W Illustrations
Calendar
Calendar Month
Call Number
Camp
CA+P Accession No.
CA+P Course
Caption
Caption on slide
Cataloged by
Cataloger
Catalog Number
Catalog Transcription
Catalogue
Catalogued By
Category

CDM 5 URL
Century
Children of Deceased
citatation_issn
Citation
citation_author
citation_conference_title
citation_date
citation_dissertation_institution
citation_dissertation_name
citation_doi
citation_firstpage
citation_inbook_title
citation_isbn
citation_issn
citation_issue
citation_journal_title
citation_keywords
citation_language
Citation_lastpage
citation_other_author
citation_publisher
citation_technical_report_institution
citation_technical_report_numbe
citation_title
citation_volume
City
Class
Classes
Classification
Click here for an immersive experience!
Click here for a 'page-flip' view
Click here for details
Click here to go to AFRC 2012 Winter Conference Materials
Click here to go to link
Click here to view photo set!
Click to Play
Clinical
Clinical Signs

Cloth Discount
Cloth ISBN
Cloth Price
Cohort Number
Collection
Collection aliases
Collection compiler
Collection Compiler
Collection contact information
Collection information
Collection Information
Collection, Is part of
Collection Name
Collection Name and Number
Collection number
Collection Number
Collection number and name
Collection Number and Name
Collection Owning Institution
Collection type
Color Illustrations
Comment
Comments
Commissioned By
Common Name
Company
Company/Business
Complete Catalog
Completion Date
Compressed file editing softwar
Compressed File Extent
Compressed file operating syste
Compressed file quality
Compressed file size
Compression
Conference
Contact information
CONTENTdm file name
CONTENTdm file name v.5
CONTENTdm file path v.5
CONTENTdm number
CONTENTdm number v.5

Context
Context URL
Contibuting Institution
Continent
Contrib. Institution
Contributing Institution
Contributing.Institution
Contributing Institututions
Contributing Organization
Contributor
Contributor Primary
Contributor Primary
Contributor Secondary
Conversion specifications
ConversionSpecifications
Conversion Specifications
Copy ID
Copyright
Copyright Date
Copyright notice
Corporate name
Corporate Name
Country
County
County Name
Coverage
Coverage spatial
Coverage-spatial
Coverage Spatial
Coverage-Spatial
Coverage - Temporal
Coverage-Temporal
Covrage-spatial
Creation Date
Creation methodology
Creation Methodology
Creator
Creator Dates
Creator Nationality
Creator of Media Recording
Creator(s)/Author(s)
Credit Line
Culture
Current address
Current Address
Curriculum
Data Description
Date
Date created
Date-Created
Date created v.5

Date Delivered
Date digital
Date.digital
Date Digital
Date-Digital
Date.Digital
Date.Earliest
Date Entered
Date issued
Date.Latest
Date modified
Date modified v.5
Date of photograph
Date of Photograph
Date.original
Date Original
Date-Original
Date.Original
Date Original DF
Date or Period
Date Scanned
DC.citation.epage
DC.citation.issue
DC.citation.spage
DC.citation.volume
DC.relation.ispartof
Death Date
Deathplace
Deceased Age
Deceased NameDegree
Degree Granting Institution
Department
Department email
Department name
Depth (3D Objects)
Desciption Anecdotal
Description
Description2
Dewey Decimal Number
Digital Archive
Digital Collection
Digital File Location
Digital Format
Digitization Specification
Digitization specifications
Digitization Specifications
Digitization Specs
Digitized by
Digitizing Device
DigitizingTechnician
Dimensions

Full NatureServe Report
Full resolution
Full Resolution 2
Full resolution folder path
Full resolution v.5
Full test
Full text
Full Text
Full text field
Full Text Field
Full Text Search
Full Transcript
Function
Funding
Further InFunding/Fellowship formation
Game Website
Gender
Genre
Genre (AAT)
Genre or Form
Genus
Geographical names
Geographical Names
Geographic coordinates
Geographic Coordinates
Geographic Location
Geo-locale
Geopolitical place
Get Media
Global Positioning System Coordina
Grade Level
Habitat Type
Has Part
Height, Framed
Height Unframed
Hidden Description
Historic address
Historic Address
Historic place name
Historic Place Name
History
Holding institution
Holding Institution
Holding.Institution
Holding Location
Hosting institution
ID
Identifier
Image Capture

Variations in title field name

Collection Alias	Local Field Name	CONTENTdm nickname
col/ehsl-wht	Title	title
col/epubs	Article Title	articl
col/fhwa	Document Title	title
col/fitz	Title	title
col/flwright-jp2	Alternative Title	latin
col/flwright-jp2	Title	title
col/GBA_Photos	Caption	title
col/GBM	Caption	title
col/gis	Title	title

Controlled Vocabularies

Type (DCMI Type vocab)

- Collection
- Dataset
- Event
- Image
- Image/MovingImage
- Image/StillImage
- InteractiveResource
- PhysicalObject
- Service
- Software
- Sound
- Text

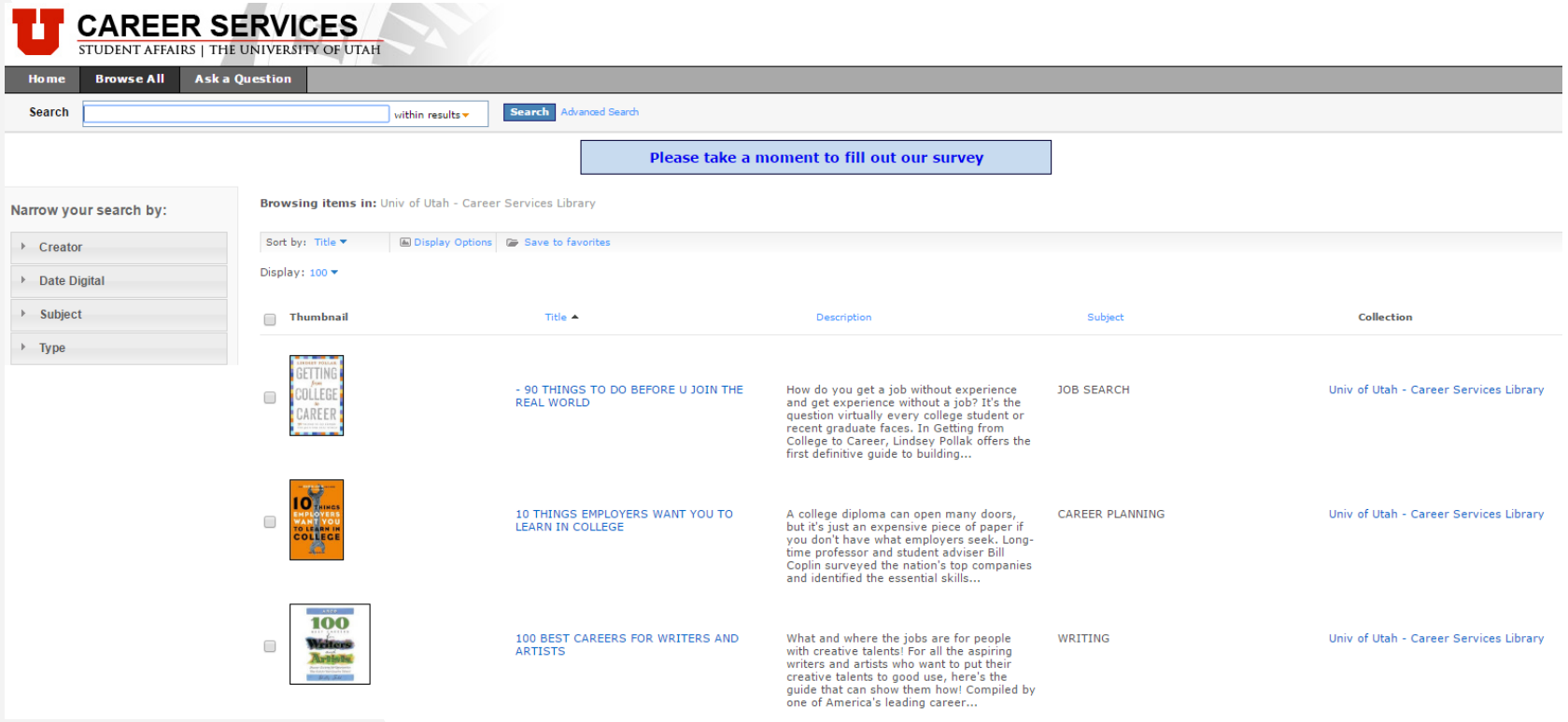
Format (Internet Media Type)

- application/msword
- application/pdf
- application/vnd.google-earth.kmz
- application/vnd.ms-powerpoint
- application/xml
- application/x-shockwave-flash
- application/zip
- audio/mpeg
- image/jpeg
- image/png
- text/html
- video/mp4

Language (ISO 639-2)

- ara
- chi
- eng
- fre
- ger
- heb
- ita
- kor
- lat
- nav
- rus
- spa

Delete Unneeded Legacy Collections



Future metadata corrections - next steps

Review 441 fields manually, noted issues (field used in only one collection, field synonyms)

Develop spreadsheet to cluster possible fixes by partner, write up metadata recommendations - ongoing process

To do:

•Explore possibilities of assessment scripts against solr to review metadata values
•Explore visualization tools

Future metadata corrections - Rights statements

40867	This material may be protected by copyright. Permission required for use in any form. For further information please contact the Multimedia Archives, Special Collections, J. Willard Marriott Library, University of Utah.
20207	http://creativecommons.org/licenses/by-nc-sa/1.0/
5134	Digital Image Copyright 2008, University of Utah. All Rights Reserved.
4325	Digital Image Copyright 2010, University of Utah. All Rights Reserved.
3597	Copyright owner unknown
3583	Digital image copyright 2005, University of Utah. All rights reserved.
3488	Digital image, copyright 2003 Uintah County Library
3405	This material is protected by copyright. Express permission is required for use in any form. For further information please contact the Multimedia Archivist, Special Collections, J. Willard Marriott Library, University of Utah.
3033	This material may be protected by copyright. Permission required for use in any form. Please contact the Multimedia Archives, Special Collections, J. Willard Marriott Library, University of Utah.
2975	This material is protected by copyright. Permission required for use in any form. For further information please contact the Multimedia Archives, Special Collections, J. Willard Marriott Library, University of Utah.
2616	Digital Image Copyright 2007, University of Utah. All Rights Reserved.

Future metadata corrections - MWDL Application Profile

title	date	description	subject	type	format	rights
National Ski Association officers and directors	January 13 - 15, 1961	Photo shows National Ski Associat	Melville, Alton Crane, 1	Image	image/jpeg	This material may be prot
Group Picture in the Orderville Ward				Image	image/jpeg	This material may be prot
Saltair Beach Pavilion and bathers, Saltair Resort.		1893-1925 (destroyed by fire)		Image	image/jpeg	This material may be prot
Interior of the Dancing Pavilion, Saltair Resort.		1893-1925 (destroyed by fire)		Image	image/jpeg	This material may be prot
Saltair Beach Pavilion, Saltair Resort.		1893-1925 (destroyed by fire)		Image	image/jpeg	This material may be prot
Group of ski tourers high in the Wasatch mountains		A group photo of ski tourers in the	Wasatch Mountains (U	Image	image/jpeg	This material may be prot
Painting of soldiers.					image/jpeg	This material may be prot
George Thompson in England.				Image	image/jpeg	This material may be prot
George Thompson with second wife Katherine.				Image	image/jpeg	This material may be prot
William Jennings Bryan at the radio station.					image/jpeg	
Nels Just with Emma Bennett Just				Image	image/jpeg	This material may be prot
Emma Bennett Just's son Fred Bennett.				Image	image/jpeg	This material may be prot
Charley and James Just				Image	image/jpeg	This material may be prot
Nels Just				Image	image/jpeg	This material may be prot
Thomas Moran				Image	image/jpeg	This material may be prot
Just family by their ranch house				Image	image/jpeg	This material may be prot
Francis and Will Just				Image	image/jpeg	This material may be prot
Emma Bennett Just at age 16.				Image	image/jpeg	This material may be prot

Future metadata corrections - Subject faceting

Topic

American newspapers	1
Antisemitism	12
Auschwitz (Concentration camp)	1
Bamberger Railroad Company	2
Bar mitzvah	1
Cantors (Judaism)	1
Conservative Judaism	1
Eastern Front (World War (1914-1918))	1
Emigration and immigration	3
Ethnic groups	1
Evacuation and relocation of Japanese Americans (United States : 1942-1945)	1

Roadmap for future development

Solphal (Front end) development

•International Image Interoperability Framework (IIIF)
•RDFa and schema.org support

SIMP Tool enhancements

•Workflow automation
•Large updates monthly, small updates as needed

Large projects

•IMLS Grants
- Newspapers in Hydra (LG-70-17-0043-17)
- Western Name Authority File (LG-72-16-0002-16)

•Born digital content
•Partner projects
•Data Repository

People

Platforms

Processes

Lessons Learned

Wild wild west of metadata wrangling

Shooting of Patrick Coughlin
https://collections.lib.utah.edu/details?id=962118

Don’t keep shoveling
while it is still snowing!

Unidentified man holding snow shovel in front of his house, (family, friend or acquaintance of Norman D. Nevills).
https://collections.lib.utah.edu/details?id=984034

Thank You!

Contact us at:

Jeremy Myntti (jeremy.myntti@utah.edu)

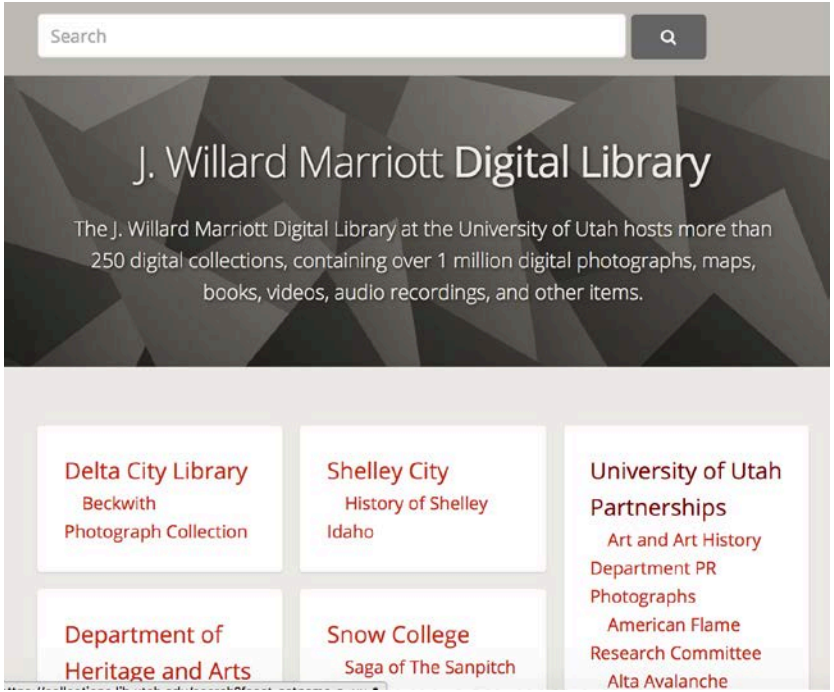
Anna Neatrour (anna.neatrour@utah.edu)

Harish Maringanti (harish.maringanti@utah.edu)

Brian McBride (brian.mcbride@utah.edu)

Alan Witkowski (alan.witkowski@utah.edu)

Matt Brunsvik (matt.brunsvik@utah.edu)



https://collections.lib.utah.edu/
https://newspapers.lib.utah.edu/